第２回ＡＩを活用したプログラム医療機器に関する専門部会

日時　　令和４年９月２６日（月）
１５：００～
開催形式　Ｗｅｂ会議

＜開会＞

○事務局（渕岡先端技術評価業務調整役）　第 2 回 AI を活用したプログラム医療機器に関する専門部会を開催させていただきます。本日はお忙しい中お集まりいただきまして、ありがとうございます。

＜委員出席状況報告及び配付資料確認等＞

○事務局（渕岡先端技術評価業務調整役）　委員の出席状況について報告申し上げます。現在、欠席と御連絡を頂いておりますのは鎮西委員、それから中岡委員が少し遅れているようです。また事前に遅れるということで御連絡いただいているのが、陣崎委員と武田委員となっております。したがいまして当委員会の 17 名の委員のうち現在 13 名に御出席いただいておりまして全委員の過半数に達しておりますので、専門部会規程第 7 条の規定に基づき、本委員会の成立を御報告といたします。

次に配付資料の確認をさせていただきます。画面に表示されております議事次第・資料目録の他、資料取扱区分表、それから資料 1～4、参考資料 1 です。資料に不足等ありましたら事務局までお願いいたします。

次に資料の取扱区分表を御覧ください。資料は内容に応じて取扱いとして厳重管理、取扱注意、その他に分類しまして、それぞれに応じた対応を取ることとしております。本日の配布資料 1 と参考資料 1 はその他に該当しまして、委員各自で適切に保管、管理、廃棄をお願いいたします。資料 2、3、4 については取扱注意のため厳重に保管いただいて、コピー等の不正、第三者への開示は御遠慮いただくようお願いいたします。

それでは、佐久間部会長、議事の進行をお願いいたします。

＜AI を活用したプログラム医療機器に関するご講演と意見交換

①「診断支援 AI 開発の評価データに関する論点および提言」

富士フイルム株式会社 メディカルシステム事業部 IT ソリューション部 統括マネージャー 成行書史氏 （ご講演および意見交換）

富士フイルム株式会社 メディカルシステム開発センター IT グループ部長 桝本潤氏（意見交換）＞

○佐久間部会長　皆様、お集まりいただきましてありがとうございます。それでは議題に移ります。8 月 10 日に開催された第 1 回専門部会ワーキンググループの参加者と概要を、資料 1 におまとめしています。論点については本日の後半で議論することといたしますが、まず本日は産業界と委員の方から講

演を頂くことになりました。まず産業界ということで本日は AI を活用したプログラム医療機器の開発の立場からの課題について、富士フイルム株式会社メディカルシステム事業部から御講演いただきます。成行様と桝本様に御参加いただいております。医機連を通して今回のデータの活用、データを作るところでの問題等について御意見を頂けないかということでお願いいたしました。それでは富士フイルムさん、よろしくお願いいたします。

○成行様　佐久間先生及び委員の先生方、このような機会を頂きありがとうございます。

本日のアジェンダですが、先生からお題を頂いております。データのあり方について、我々のほうで諸々課題意識等を説明させていただきます。我々富士フイルムはそもそもどんな AI 及び SaMD をやっているのかといいますと、内視鏡から始まり X 線及び CT、MR を含めた様々な画像診断システムを開発して案連するおりまして、さらに得られた医療情報に関する SaMD 及び AI 機能等の開発及び社会実装を進めております。こういう背景なので、本日は主に、画像診断の領域に関したコメントになるかと思います。具体的には、AI 技術を活用したプログラム医療機器において、予防から予後へのバリューチェーンの中で、主に検査支援、診断支援から治療方針支援といった辺りのプログラム医療機器がスコープとなります。

では早速ですが、評価データに求められる条件について。当社は CT の肺結節の検出 CAD を承認申請して認可いただきましたが、その際の評価として単体性能評価および読影試験を行い、その結果を申請書に添付しております。試験にあたっては、カンニングができないように、データの分離に関して内部でしっかり配慮をしております。例えば単体性能・読影評価用のデータは、開発に使った学習データ・検証データとは別の施設から入手するであったり、読影実験においては読影医の所属施設と重複しない施設から症例画像を入手するということであったり、バリエーションを考慮して複数の医療機関から入手、医療機関のカテゴリー・規模などもバラエティに富むように努力をしております。上記は、PMDA とのやりとりにおいては、一般的（標準的）な取組かと思います。

次に、どのようなデータを収集しているかですが、肺結節のケースでは、結節のサイズ・場所などについて網羅的に収集しています。どこまで細かく分類するかによってバリエーションが結構広がるということです。

肺結節の種類（Solid、Part Solid、Pure GGN といった所見上のバリエーション）に加えて、サイズ、位置などでバリエーション全体は 84 通りにもなります。さらに CT 画像なので、装置のバリエーションがあり、それも掛け算になります。これに対して、後ほど整理して述べますが、アルゴリズム特性から必ずしも確認が必要ないものも含まれているのではないか、更にはできるだけ広いバリエーションを評価するというよりも、工学的及び医学的な観点から確認が必要なバリエーションに絞って収集する方が望ましいのではないかという見解を持っております。

現状の課題意識としましては、やはり負荷が大きいこと。下に３つほど書いておりますが、集めたいと思ってもなかなか集まらないものも当然あります。実務的には、契約や先生のご協力を頂くための手続きもなかなか大変です。もう１つは人材の話です。我々は医療機器メーカーですので医療機器及び臨床を理解している人材をしっかり社内で持つ努力が必要ですが、複数プロジェクトのためのリソースであったり、昨今 SaMD は関心が高く、人材の獲得競争状態にあるため、開発及び社会実装を加速しようとしても、人材が律速になるケースが結構あります。

網羅的なバリエーションの例では臨床側の話をしましたが、入力画像の条件で前組合せは 1,680 通りとなって、真面目にやろうとしてもそもそも無理といったところは常識的に考えてもあるかと思います。

これらの状況において、できる限りやるという話ではなく、もう少し観点を整理し、バリエーションを絞る方向に考えた方が良いのではないかというのが、我々の提言になります。率直に言いますと、現状は PMDA から明確な指針が整理されて発出されているわけではなくて PMDA の審査官の方と議論しながらどういう評価をすれば承認いただけるかといった観点から、手探りで妥当な評価条件を提案しており、結局は一般的に認知されている装置の仕様項目とか撮影パラメータ、又は疾患・所見のバリエーションなどを中心に、ケースバイケースで議論しているのが現状です。

今の SaMD の状況を踏まえて PMDA 内でも検討が重ねられていると思いますが、ここまで何件かを審査・承認いただいてきた中での我々開発側からの提案としては、技術・アルゴリズム特性に基づいたバリエーションを定義した方が合理的ではないかということです。例えば、病変の場所というパラメータがありますが、臨床的な場所のバリエーションをそのまま評価する必要性がないケースもあると考えています。用いた AI 技術

の設計・特性から見て同じカテゴリーに入ると整理することもできる。実務的には PMDA 側での臨床知見を持った審査官及び医師に近い方々から、臨床的観点からこれこれのバリエーションを評価すべきではないかという提案をいただいたら、設計側からは技術的に整理したバリエーションのこのカテゴリーに含まれるので、新しいパラメータ・バリエーションとして広げる必要はないのではないと思われるといった議論を通して、技術に基づいた評価すべき最適（ミニマムな）バリエーションを定義できるのではないかということ、、我々が考えていることです。

それ以外にも、CT 及び MR など装置のバリエーションも装置の特性によってメーカー依存性が少ないパラメーターもあるといったことや、バリエーションとして複数評価するよりもワーストケースを特定して評価すれば担保できるような項目もあるとも考えております。

次のトピックは、読影試験でのバイアス回避、有病率の問題で、開発側にとっては大きい問題です。読影試験に関する PMDA とのプロトコル相談において、実際に使用を想定する有病率、例えば健診なのか、診療シーンなのかといったところを特定し、有病率をしっかり仕様に合わせて評価するようにという指摘がありますし、標榜にもそのような内容を特定し記載するような方向で指示をいただくことがあります。例えば健診は AI 診断支援技術のような SaMD の使用に関する関心が高いと思いますが、健診での有病率の再現は、開発側にとって負荷が大きく、現実的には無理だと思います。0.何％というようなケースもあると思います。どういう標榜をしたければどういう試験をすればよいか、現実的な視点に立って枠組みを示していただきたいと思っております。例えば、現実的に事前に評価が難しいものに関しては、市販後の活動を通じて製品の安全性及び有効性をフォローアップしていくようなリアルワールドデータ的な取組だったり、現在のリバランス通知の考え方を拡張又は応用するような制度といったものも議論、検討の余地があるのではないかと考えております。

少し話が変わりまして、①の評価データに求められる条件に関するトピックではないかもしれませんが、結構直面する課題として、熟練医・専門医を超えるような AI の評価に関する論点について、説明させて下さい。例えば事例 1)は、細かい話は技術的な所を文章で書いて分かりづらいとは思うのですが、要は検出 CAD で熟練医、専門医を超えるようなレベルの検出性能を達成する AI について、技術的に開発できそうな手応えがあり

ます。また事例 2)は内視鏡鑑別 AI の例で、現行品はあくまで先生が内視鏡画像から判断したものを正解としているのですが、例えば、病理画像を正解としたような内視鏡画像の鑑別 AI（CADx）というようなものを考えると、実際に専門医の検出能を超えるような性能が実用可能になってくるかと思っています。

課題①としましては、例えば熟練医でも対象画像からは視認が非常に困難な病変候補を GS として評価をしてよいのか。そもそも「真」の正解データはどのように定義したらよいのか、正解を定量的にどのように決めればよいのかといった課題があると思っております。

一方、課題②に書くように、要は医師が例えば対象画像からだけだと読み取れないようなものをこれが検出されましたと提案するようなものであったり、鑑別結果を示すようなものを最終的には医師の責任で使っていただくことになるわけですが、そもそも商品として受け入れられるのかという課題でもあるとは認識しております。将来技術において課題となりそうなトピックとして提示させていただきました。厚労省のほうからは熟練医、専門医を超える AI を出してくれれば保険点数を付けたいのだけれども、というようなコメントを頂きますが、この辺りが整理できないとそもそも保険にまで話が到達しませんし、こういった AI はそもそも商品化できるのかという課題意識も含め、ここで紹介させていただきました。

こちら（SaMD に関する認証基準の議論）は参考までに添付いたしました。

2 つ目のトピックのデータの再利用のあり方ですが、ここは特に新しい観点はありません。評価データの再利用はある程度しないと現実的な負荷という観点で大変ということもあり、あまり厳しくする必要がないのではないかといった観点でのコメントをさせていただいております。IDATEN で想定されるようなケース（カンニングがないという前提ですが）、データを数回再利用するような部分で過学習のリスクがないのではないかと思いますし、実用的な運用としてしっかりカンニングさせないための預託の仕組み（独立した試験機関等々の仕組み）も考えられると思います。あとは、汎化性能に対して影響するような条件が何か、どういったところに留意するべきかといった基準を示していただけると非常にありがたいと思っております。

3 つ目のトピックです。あまり議論に資する観点を御提示できませんが、

Adaptive AIというのは本当に現実的なのかというところがあります。段階的には、まずは（Adaptive ではないかもしれないのですが）パラメーターを選ばせるような AI を提供するのがステップとしては先にくるのではないかと思っております。また学習データは追加しても、学習させるフレームの変更は想定していないケース（我々がしっかり開発・管理できるもの）が想定可能な範囲かと思っております。

あと実際の運用を想定した場合に考慮すべき点としては、企業が責任が持てるのがどこまでか？、見落とし防止の AI であって感度が上がる方向であれば許容するという考え方など。開発側と使用側の責任範囲の明確化といったところで、今は最終的に医師の責任で使用するといった基準が示されていますが、こちらも改めて整理をする必要があるのかなと考えております。

論点としてはここまでとなります。最後アペンディックスとして添付しておりますのは、Adaptive AIという観点でオンラインで実際にアノテーションしたり学習をしたりといった仕組みとして使いうるクラウドのサービス（これは研究開発用ですが）を開発してちょうど提供を始めたところという御紹介になります。AI 開発プロジェクト関連のアノテーションから AI の推論実行までをワンストップでできるような仕組みは、弊社以外も提供していますが、医療に特化した部分でもこういったサービスの社会実装が始まっているという例の御紹介になります。御清聴ありがとうございました。

○佐久間部会長　ありがとうございました。それでは委員の先生方から何か御質問ございますでしょうか。

今日頂いたスライドの 6 ページ目の所などに、工学及び医学的な観点からバリエーションを絞ったほうがよいのではないか。これは可能かどうかということなのですが、少し具体例としては技術的にどういうことがあるとこれはできるということになるのでしょうか。このような可能性を検討しうるような場面とはどういう状況なのかと考えられていますか、少し具体的に教えていただければと思います。

○桝本様　　そうですね、例えば肺結節ですと医学的観点は分かりやすいですよね。存在しない所に発生する病気というのはなかなか無いので、それはできればバリエーションから外したいというのと、あと工学的観点で言うと、例えば肺結節の孤立性成分がどこに発生しようが恐らくそれの抽出に関しては一定のバリエーションさえあれば網羅的なバリエーションは多分

必要ないと考えているのですが、今回の肺結節の承認申請では、PMDA 側からは、各肺の各領域に対して孤立性成分があるようなパターン網羅的に準備するとか、実臨床の画像で非常に稀にしか発生しない場所もあるのですがここは網羅性は要らないかなとか。

○佐久間部会長　途中で失礼。質問の意図なのですが、必要と思う意図と、技術的に、あるいは科学的な側面からそれがなぜ可能になるのかということをご説明いただくとその議論につながるのですけれども。

○桝本様　　分かりました。多分我々が根拠としているのは何かというと、学習データの分布です。そのパターンに関しては十分に検知できるという自信を持って、そこの網羅的なバリエーションは要らないだろうという主張をしたいのです。ですので学習データのばらつき等を鑑みた上で、そのバリエーションは十分網羅されているだろうと判断していただければ、比較的やりやすいのかなと思っています。

○佐久間部会長　なるほど。学習データの分布が一定の特徴を持っており，それを仮定した状況では，一定の説明が可能であろうということであるかと思いますではそこも若干考えなければいけないところもあるとは思うのですが、それをどのように保証したらよいかということを考えていらっしゃるということですね。分かりました。

○桝本様　　例えば分かりやすいところで言うと、学習データではある程度バリエーションを限定的に強引に集めたデータですので、一定時間に時系列順に集めてなどの条件がなければ、我々でも網羅的に学習データを集めることが可能で、そのデータを使って学習させているのですが、PMDA の承認申請の評価用に集めていると、先ほど成行が言ったように実際の臨床現場に沿った発生頻度でのデータ収集が求められるのです。そうすると、その収集方法で網羅的データというのをなかなか集められないというのが現実問題としてあるというのと、学習時は、例えば肺結節ですと、コンピュータで擬似的にその場所に癌のようなものを作って学習させることもよくあるのです。そうすると孤立性成分として存在しない、なかなか臨床データが手に入らないところのデータでも擬似的には作れるので、そういったところはできれば学習データの網羅性で、実データのほうはある程度サンプリングした形で認めていただけないかなという、分かりやすく言うとそういうデータになります。

○佐久間部会長　今のはインシリコと言うかそのラーニングで人工的な学習データを作ってデータを水増しするということが実はあり得るのでしょうが、それは

どこまで現実をフォローできているかという対応の部分をどうするかということですね。この点、委員の先生方から何かコメントございますか。

○伊藤副部会長　非常に面白い話だったのですが、今ちょうど話題になっていた、偽の結節をいろいろな肺の領域に乗せて学習させる話でね。その結節に関しては、いくつかの性質やサイズの違う結節を作成し、学習させたら効率的だと思うのです。むしろ CT の数少ないものを探すよりはよほど学習効率が良いと言うか、実際にそこの AI のモデルを作るのに非常に良いと思うのですが、1 つとても気になったことがあります。結局網羅的な肺結節の CAD のこういったものを作ったときに、CAD、ｅもｘも作ると思うのですが、結局、臨床的な観点か、一番右の所に医学的な観点と書いてあるのですが、小さい 2mm とか 3mm の結局 GGN などの疑いを見つけても我々は何もしないのです。ですから多分医学的な部分で、どういったところにディシジョンの変化し得るものがあるのかというところを医学の専門家とどれぐらいディスカッションしたのかなという部分が少し気になったのです。

正直言うと GGN のようなものはほくろのような良性の性質を多分に含むものなので、もう 5mm 以下のほくろなどは多分放っておくわけなのです。ですのでやはりこれはもう ROC カーブで 9 割の精度でがんですよといった場合はこれはもう一大事で早く治療しなければいけないので、その辺の医学的などこの部分をターゲットに持っていくかというのも今のお話を聞くと結構重要かなと思いました。

○佐久間部会長　ありがとうございます。他の観点で何か。

○桝本様　今の観点はとても良い御指摘だと思っています。我々も正解データのボーダーラインをどこに引くかというのはとても悩ましくて、臨床的に意味があるものを見つけるという医師の主観が入るボーダーラインはもちろん、例えば 5mm 以上とか 1cm 以上という明確な定義をして、先生方に正解を作っていただいても、ボーダーラインが人によってブレるのです。そこの正解の付け方によって AI の性能評価が混乱してしまうというのがあって、実はなかなか医師の主観が入る微妙なラインでのボーダーを引くというのが難しいという現状があります。我々としてはどちらかというと、今回の肺結節 CAD の例のように、臨床的にその後フォローするかどうかというのは先生方に依存するとして、ある程度目に見えるような結節を全部検出するような AI をまず最初に作らせていただきました。そうすると見えるか見えないかという判断だけなので、比較的正解データがばらつかなくて性能評価がしやすかったというのが現状、実際問題とし

てあります。一応、参考までです。

○佐久間部会長　ありがとうございます。他に委員の先生方、何かございますか。

○村垣委員　　この話題とは少し離れるのですが、企業戦略として、例えば肺結節のCADみたいにそれ自体で単体のアプリとかソフトとして診断を求めてAIとしてやる場合と、例えば画質の改善だとかセグメンテーションを例えば会社の中のCTとかMRIのソフトの中にバンドルしてやるAIという２つの方法があると思うのです。後者の方法で、単体でそれ自体で薬事を取ろうとしているものとバンドルしているものとで何かAIの基準が違ってきたりとか、その辺のところはどうなのですか。

○成行様　　そうですね、基本的に開発する時点のアルゴリズムとしては、そこはなるべくユニバーサルに使えるようなものとして作っています。

○桝本様　　でも現段階ではおそらくそこにあまり差を付けるつもりはなくて、今後まずは汎用的に使えるものを作らせていただいた上で、その後は例えば再構成する画像の前の情報を使って性能を上げるとかということで、差別化要素的な観点で自社向けの専用AIというのを作る可能性はあると思うのですが。

○村垣委員　　全体的な数的に、そのような単体の承認を目指してやるAIと、自社内での性能向上のためのAIというものは開発割合的にはどうですか。

○成行様　　個々の会社の開発の取組みによるのですが、我々の取組としてはやはりいわゆるPACSとかワークステーション側で展開するような汎用化しているAIのほうが圧倒的に多いです。

○村垣委員　　その辺のところも是非、同じ基準なのか基準が違ってくるのか、どこを最低限求めるかなどを、教えていただければと思います。

○佐久間部会長　今の議論はバリエーションのところで機種間の差をどう考えるかとか、撮影条件の環境をどうするかというところをコントロールできるかどうか、そこでバリエーションを落とすということの戦略としてはそれは機器に固有にしてしまうという考え方はあると思うのです。

○成行様　　そうですね。実際になかなかデータが集まらなかった場合には、機種を絞って標榜してまず上市しているものもあります。

○佐久間部会長　一方でそれがやはり現実問題としてはスライス厚の違いなどによって、どうも性能が違いそうだということも出てきているようなので、その辺りの知見が蓄積されてくると削減の方向には行くのかもしれません。少しずつそういう事例を集めてくることも必要なのかもしれないですね。そこら辺はきちんと皆さんで情報を共有してうまくそこを乗り切れるよ

うな体制を作っていくのは結構重要だと思います。他に委員の先生方、よろしくお願いします。

○清水委員　15 枚目のスライドで確認させていただきたい所があります。Adaptive AI についてちょっと興味があります。上の 2 行のまず 1 行目の「追加学習ではなく、パラメーターを」とありますが、これは ROC 上の閾値を変えるというお話でしたか。

○成行様　そうですね。メーカーが完全に管理した状態でということなので Adaptive ではないものになります。

○清水委員　関連施設で 2 つ目に、「モデルの変更は想定しえない。あくまで学習データの追加による。」とありますが、これは学習データの追加によって何かが変わるのですか。変わることを想定されているのですか。

○成行様　ここに書いているのは、学習モデルまで大きく変えるような意味での変更は想定していないということです。

○清水委員　ネットワークを全く違うものに変えるとか。

○成行様　そうですね。

○清水委員　ネットワークの個数とか。

○成行様　例えばそういうことですね。

○清水委員　そういうことは想定せずに、学習データの追加による再学習は考えているということですか。

○成行様　そうですね。どのような性能に変化してしまうか想像がつかないようなケースというのは、メーカーとしては、現実的とは思えないという意味でコメントさせていただきました。

○清水委員　ここは一番最後の責任問題と強く絡んでモデルの変更による精度の変化というのは予想しづらいので、学習データの追加による再学習ぐらいまでならメーカーとして責任が持てるような開発ができるだろうというようなお考えということですね。

○成行様　そうですね。学習データ、追加させるデータが何パーセントまでといったような制限など。

○桝本様　1 行目と 2 行目の境は相当大きいと思っていて、1 行目の考えは比較的短期間で我々も製品化したいと思っています。2 行目の学習データの追加によるモデルのほうは多分やるとしたらその範疇内だと思うのですが、実際にやれるかというと現段階でやれる自信は全くないというか、恐らく現地での学習の方法によりますが悪影響のほうを抑えきれないと思っているので、メーカーとしては今のところ手を出せないとは思っていま

す。

○清水委員　悪影響のときには止めるということも書いてあるのですが、そういうような運用でも、何か具体的な御心配というのはあるのですか。

○成行様　悪影響の判断のしようがないというところが、一番の問題だと思うのです。

○清水委員　例えば判断の方法としては、市販前のデータ、承認を受けたときのデータなどを使って評価をする。それでその評価結果が承認を受けたものより悪くなれば止めるというような、単純な方法もあると思うのですが。

○成行様　そうですね、少なくとも悪影響はないだろうと見切れているようなケースが仮に想定できればあり得るのかもしれないです。しかし、学習データの追加によって、元の評価データセットで評価し切れていない部分で悪影響が出る可能性というのは、考え得るかなと思っております。

○清水委員　なるほど。例えば破滅的忘却なども、心配されているのですか。

○成行様　はい。

○清水委員　分かりました。どうもありがとうございます。

○森委員　お伺いしたいのです。まず、Adaptive AI は多分やらないとは思うのですが、こういったときに性能の一貫性とか傾向をどうやって保っていくか。例えばある製品がバージョン 1 だとこれで、多分 Adaptive ではなくて、市販後学習でユーザーサイトで学習ではなくて、企業サイドでも学習を指した場合でも同じようなことが起きると思います。どういうことかと言うと、いわゆる性能的には結構上がっているのだけれど、全然以前とは違うよねといった形とかが出てくると思うのです。ですからユーザーは使っていった上で機械の傾向というものに結構慣れてくるので、大体どのような傾向があるからこれはトゥルーだと言っているけれどもネガティブではなくて、ユーザー自身も大体の傾向を学習していくので。それが変わらないようにするために何か工夫をされるのかどうかということが 1 つ目。

2 つ目が文書で頂いていたほうで、J-MID のデータベースをどのように活用されるのか。御存じのように NII 側にもいるので、2 つ目の質問が一番聞きたいのですが。お願いいたします。

○桝本様　まず我々のほうで何がやれるかというと、メーカー側は再学習は当然ながら性能向上のためのバージョンアップのためにやるのですが、担保している方法としては、我々は承認に使ったデータだけでしたら、今、清水先生がおっしゃったようにやはり数が少ないので、今、森先生がおっ

しゃったように、ばらつきが違うパターンで、性能は上がっているけれども何か傾向が違うなと感じることは、我々もあるのです。ですので手元で持っているその数十倍から数百倍のデータで統計的な情報を見ています。

そこでばらつきなども前回よりも明らかに良いよねと分かってもらえるような状況を確認した上で改良のほうは進めているので、そこは問題ないかなと思っています。

もう１つ JMID の話はおっしゃられるようにすごく難しくて、我々としては今手元に持っている数十倍、数百倍と言っているのが、JMID を使えば、おそらく数千倍、数万倍となる可能性があり、画像のバリエーションも相当豊富ですので、よりそこの安定性に寄与できるのではないかなと思っています。

○森委員　なるほど。今、ジェミとか我々転送を受けているほうが現在 3.6 億枚入っているので、ジェミとか来ていないデータもあるかもしれませんが大体それぐらいのデータがあるので、これをどうやって利活用するかというのは、この PMDA の会議とは別になりますがどこかでまたできると良いかと思っていますので、よろしくお願いいたします。

○桝本様　ありがとうございました。JMID のデータは、今は学習に使うのは非常に困難ですので、評価データとしての大規模データとして使わせていただいています。

○森委員　ありがとうございます。

○佐久間部会長　何か委員の先生方、ございますでしょうか。

○佐々木委員　病理学会として実は病理診断に関する AI のプログラムについて、保険収載を目的にいろいろ進めてきたのですが、その際にいわゆる病理医が 1 人しかいない病院が非常に多くて、そういう病院ではダブルチェックをせずに病理診断が出ている。ダブルチェックをするのに正にスライドにもありましたが、病理診断管理加算 3 というのを付けてはどうかと。その場合ダブルチェックをさせるだけが目的なので、別に病理医の診断性能を超える必要はなくて、むしろ拾いすぎをどうやって抑えるかが課題になったことがありました。

逆に拾いすぎると、最終的には我々はこの AI のプログラムをどう使うかというと、一人病理医が 1 回診断をして、それでダブルチェックを AI がやる。AI がもう 1 回見てくださいねとアラートを鳴らしたものだけもう１回見るという活用の仕方を考えて、そういうシステムを入れてダブル

チェックを実行しているところに病理診断管理加算の 3。病理診断管理加算の場合には一人病理医の場合には 1,200 円、それから 2 人以上いてダブルチェックの体制がある所が 3,200 円と加算が付いているのです。ただ病理医の数が少ないので、1 人しかいない所は 2 人目の病理医の代わりを AI がやるということで病理診断管理加算の 3 と付けて 2,200 円というお金を申請したのです。どういう目的で AI を使うかということをまず最初に決めないと、あまりに低すぎる AI を作っても意味がないのかなと思ってお話を聞いていました。ですのでそういう使い方をするのであれば医師の性能を超える必要はなくて。ですので厚労省側も言っていることは理解できると言ってくださったということなので、やはり保険点数が付かないとなかなか使えないと思うので、そこで話を進めていただければということで、参考までにお話をしました。

○成行様　　そうですか、ありがとうございます。理解いたしました。

○佐久間部会長　他によろしいでしょうか。皆さん活発な御議論、ありがとうございました。

これで富士フイルムのお二方、申し訳ないのですがこれで Web 会議は退席していただきます。今日は貴重な御意見ありがとうございました。うまく議論の中に反映させていきたいと思いますので、また今後ともよろしくお願いします。

○成行様　　よろしくお願いいたします。ありがとうございました。失礼します。

○桝本様　　失礼します。

○佐久間部会長　今のところは多分今日森先生がおっしゃったことで後での議論になると思うのですけれども、使っている側がその AI の特性に慣れてしまって学習した結果として変わったことによって、実は違う医学的な判断をしてしまうというのがあって、それも実はある一定の確かにバイアスのところに出てくるのです。

それからあとはデータがいわゆる学習のデータの分布に十分なものがあればよいのではないかという議論、これについて確かにそう言えるところもあるのだけれども、少しまた深く議論をしないといけないこともあると思うので、その辺りをまた後ほど議論することにします。

あとは 3 番目に大規模データのところをどうするか、今日後半の話にちょっと関連しますので、森先生、別にここで少し議論してもよいことかと思いますので、よろしくお願いします。

○森委員　　分かりました。

＜AI を活用したプログラム医療機器に関するご講演と意見交換

②「プログラム医療機器の市販後再学習における性能評価」（清水委員）＞

○佐久間部会長　それでは次に清水先生から、医療機器の市販後再学習における性能評価について御講演を頂きます。ちょうど今日 Adaptive AI のところで入れたことがどう変わるかといったこと、それから過学習のこと、そういうことについての最近の知見について、お話いただけるかと思います。それでは清水先生、お願いします。

○清水委員　それではプログラム医療機器の市販後再学習における性能評価ということで、今日はこの 3 つのお話をさせていただきます。配付した資料から一部変わっているところもありますが、こちらの資料を使って説明します。

まず 1 つ目がそもそも評価が必要なケースはどんな場合であるか。それから評価法としてどういうものが考えられるか。3 つ目に評価時のリスクについて、話をいたします。

まず最初に評価が必要なケース。特に顕著な例を持って来ました。それが市販前後で特徴が大きく異なる場合、いわゆるドメインシフトが起こっている場合、再学習により性能が変化し、場合によっては破滅的忘却と呼ばれるリスクもあります。例えばこのように大きく異なる場合です。

それからここで考えないといけないこととしては、市販前(承認時)、それから市販後の Test 用 DS、DS というのはデータセットの略です。正式名称を書き忘れていますが、データセットによる評価が必要。これは両方必要なのか、片方でよいのか。片方というのは市販前承認時はこれは欠かせないだろうと思っていますので、そちらだけでよいのかということです。恐らくこういうことも議論になると思います。

その次、では評価法はどういうものがあるのか。これは非劣性や同等性の評価があると思います。評価する場合はまず主要評価項目、そのプログラムの目的に応じて例えば AUC を使います。Accuracy を使います。セグメンテーションであれば DICE を使います。リグレッション（回帰）の場合であれば回帰誤差などが考えられると思います。こういうものを選択する必要があります。

それから非劣性で評価しますか、あるいは同等性で評価しますかということを決める必要があります。同等性の場合には上限を考えないといけないということです。これは以前も少し議論があり、性能がどんどん良くなればよいのではないかという議論もあったのですが、以前別の所で

議論したときには性能があまりにも上がり過ぎると医者のプログラム医療機器の使い方が変わってしまうのではないか。当初の性能、例えば 70％で承認されて市販された後にどんどん性能が上がって 90％とか 99％となったときに、医者とそのプログラム医療機器の関係が変わってしまうのではないかということも懸念としてありました。その場合には上限も設ける必要があるかと思います。

それから許容範囲。これは非劣性を考えるときには非劣性のマージン、同等性の場合は同等マージンを決める必要がありますが、どういう数字を使うかを決める必要があります。

それから次のステップとしては、再学習後の性能。例えば市販前承認時の Test 用 DS により評価し信頼区間推定をしますが、よくある 95％でやるかどうかという話です。例えば市販前、これは縦軸が Accuracy で左側が市販前のパフォーマンスを表していると思ってください。この青い実線が例えば非劣性のマージンだとしますと、この一点鎖線よりも信頼区間がはみ出た、例えば 2 番、5 番というのがこれは劣性ということでパスしないということになって、1、3、4 がパスするということになります。同等性の場合には上にもマージンが設定されて、それと信頼区間との関係を使って評価することになるかと思います。

この辺りは医療機器だけでなく薬等の承認でもよく行われていることとなりますので、その辺りが参考になるのではないかと思います。

それから今日の本題なのですが、評価時のリスク、テストデータ再利用のリスクについて話をします。同一の Test 用 DS を使った評価を繰り返し行う際に Test 用 DS の結果を分類器等の設計に反映させることで、評価値にバイアスが付加されるリスクです。

どういう場合があるかと言うと、1つ目が複数のモデルを同一の Test 用 DS を用いて評価し最良のモデルを選択するということ。これは何度もこの会議でも議論があったと思います。それからもう１つ、例えば１つのモデルを同一の Test 用 DS を用いて評価し先ほどの統計的評価をして劣性だった場合には、再学習。再学習の方法としては、学習データを変えたり学習アルゴリズムを変更。その後同一の Test 用 DS を用いて評価。これを非劣性になるまで繰り返す。

開発者側としてはこういうことをやりたくなる気持ちはよく分かるのですが、これもやはり高リスクの例、バイアスが付加される例、本当の事が分からなくなってしまう例かと思います。

リスク低減の補正の方法ですが、1 つが Test 用 DS のサイズを十分大きくする。先ほど NII のデータベースの話がありましたけれども、ああいうものを活用して大きくする。2 つ目が Test 用 DS を毎回変更。完全に新しいデータにするのが理想的なのですが、例えば Test 用 DS に摂動(ノイズ)などを付加させるとか比較的大きめの Test 用 DS の集合からリサンプリングをすることによって新しいデータのサブセットを作って、それで評価する。それから新しいデータを一部に追加などが考えられます。

3 つ目が後で紹介する Test 用 DS による評価値を補正する方法で、Dwork、あるいは Gossmann らの方法に従って補正をするという方法が考えられると思います。この Gossmann の方法について少し紹介します。これは FDA のスライドを持ってきているのですが、もともとは Gossmann らの仕事が元になっています。

理想的な方法がまずここに書いてあります。トレーニングデータセットがあって、トレーニングデータはトレーニング用のセットとチューニングセット、例えばバリデーションセット。あるいは手元にあるテストセットもここに入るかと思いますが、こういうものを使って学習をさせたりモデルを選択したりします。

正しくそのシステムの本当の性能を知ろうとすれば、ここにしっかりとした壁を築いて普段はテストデータ、アクセスできない独立なテストデータを用意します。評価するときだけファイアーウォールを通してこのテストデータにアクセスして FIX された学習用モデルの性能を評価して、それをテストの性能とする。これがこれまで行われてきた正しい評価の方法となります。

このスライドで指摘しているこの赤い枠の所に注目していただければと思いますが、そもそもテストセットはそんなに潤沢に用意できないので、何度もテストセットにアクセスするような場合にはいわゆるテストセットがトレーニングセットの一部分になるリスクがあると言っています。特にテストの結果を見てアルゴリズムを変えるとか先ほどのモデルを選択するとか、そういうことをするとせっかくのテストデータが汚染されてしまってトレーニングデータの一部になってしまいます。

正しい分布からいつもフレッシュなデータを集めることができれば、そのフレッシュなデータをテストデータを見ながら設計したものに適用すると実際には性能が大きく低下しているということが判明してしまう。これがリスクだと言っています。

この下にあるDworkの仕事というのは「サイエンス」に掲載された、スレッシュ・ホールドアウトと呼ばれる方法でこれを緩和する方法です。Gossmannらはこの方法をSaMDの連続学習に拡張しました。そこで連続して学習するプロセスでテストデータの結果を見ながら、機械学習のアルゴリズムを変えていくと何が起こるかを見ています。

注目すべきグラフは、この赤い線とこの少し薄いオレンジの線となります。赤い線がTrue performanceと書いてありますけれども、毎回フレッシュなデータを用意して性能を評価した、本当の性能です。

これに対してテストデータを見ながらアルゴリズムをどんどん変えてしまうと、バイアスが掛かって本当の性能よりも見掛けがかなり良くなってしまいます。右側がそれを補正した方法で、補正した結果がこの紫色の線となっています。かなり実際の性能に近づいているのが分かると思います。

これがその論文なのですが、飛ばします。これがそのアルゴリズムとなります。このアルゴリズムではAUCを手掛かりに、テストデータの情報が漏洩しているかどうかを評価しています。テストデータに対するAUCと、トレーニングデータに対するAUC、ROCカーブの下面積ですね、パフォーマンスを表す性能。テストとトレーニングのAUCの差にηを加えた値が、このＴハットよりも大きければというものなのですけれども、直観的にはこの２つの性能が大きく離れている場合ですね、テストに適用したときの性能とトレーニングに適用したときの性能が大きく違えば、これはテストデータの情報が洩れてしまっていると考えてテストデータの性能にξ(グザイ)という、これはラプラス分布に基づくノイズなのですが、これを加えます。これを加えた全体の結果を使ってアルゴリズムを選択すれば、先ほどのテストデータを適用したときの見掛けが非常に良くなって本当の性能と乖離してしまうというリスクを防ぐことができる方法となります。

この方法はもともと差分プライバシーの技術を応用しています。特にこれはラプラス分布ですから、ラプラスメカニズムと呼ばれている方法論に基づいています。このメカニズムを使えば、例えばε(イプシロン)差分プライバシーを保証することができる。別の言い方をしますと情報漏洩のリスクを確率的にきちんと担保することができるというメリットがあるので、そういう数学的な背景もある方法となっています。

これがGossmannらのペーパーで紹介された方法で、これは幾つかの分

類器に適用してこの赤色の本当の性能と見掛けの性能が何もしないと乖離するのですが、先ほどのアルゴリズムで補正する、ノイズを加えながらアルゴリズムを更新するということをやると、このように本当の性能に近づきます。

他の分類器、例えばロジスティック回帰、それに少し条件としてレギュラライザー（正則化項）を付けた場合の結果とか、サポートベクターを使った結果とか、ランダムフォレストとかアダブースト。他の分類器でもうまく補正できますよと言っています。

この論文ももちろんよいのですが、実際にある分類器を設計するときに頻繁にテストデータにアクセスしながらその分類器を設計するというのは、さすがに考えづらい、なかなかそこまでやらないと思います。それと今我々がこの会議でも出ている、気にしている状態というのは、例えば先ほど申し上げました複数のモデルを同一の Test 用 DS を用いて評価して最良のモデルを選択、あるいは１つのモデルを同一の Test 用 DS を用いて評価し駄目だった場合は何度も適用してうまくいくまで反復するという、こういう枠組みのときにどうなるかというのを評価したいと考えています。今日は間に合わなかったのですが、これからこういうことについても実際に評価していこうと思います。

これは最後のスライドになるのですがその前までのスライドとは少し論点が違って、先ほどまではバイアスが怖いという話をしたのですが、実際にはそんなに怖くないこともありますよということを示したものです。

Kaggle と呼ばれる世界的にいろいろなタイプのデータが集められてそれをコンペティション、チャレンジ等で使っているのですが、そこで頻繁に行われるコンペティションでパブリックとプライベートの２つのデータベースがあります。パブリックのデータベースというのは、コンテスト本番までに開発者が頻繁にアクセスすることができる性能です。それがいろいろなコンペティションの場合についてグラフがありますが、横軸がその Public accuracy を示しています。

Public のデータセットというのは、全体の 75%のデータとなっています。縦軸がコンテスト本番に使用する Private のデータセット、残りの 25%に適用したときの結果です。バイアスがあればこの赤い線で示した 45 度からずれるはずなのですが、これは幾つかのコンペティションの結果を見ても広がりがあるケースがあるのでこれは少し注意しないといけないと思うのですが、大体 45 度の線の上に乗っているのが分かると思いま

す。

ただし上位の 10%を拡大してみると、少し事情が違うケースもあります。バイアスが少し心配になるようなケースです。Public のコンテスト本番までに頻繁にアクセスした性能のほうがコンテスト本番の性能よりも良くなっているという例もあります。

こういうアルゴリズムが何をしているかというのは気を付けないといけないですし、これが正にバイアスの話なのでこういうことが起こり得るというのは Kaggle のデータベースでも確認されていますが、普通に使っていればそれほど心配するケースではないということもこのデータから見えてきます。これを踏まえて実際にいろいろな検討をする必要があるのではないかと考えています。

○佐久間部会長　どうもありがとうございます。それではここで御質問があればお願いいたします。先生、最後のデータのトップ 10%というのはどういう意味ですか。

○清水委員　　これは Public accuracy の上位 10%ということです。

○佐久間部会長　良い性能を上げているというところですか。

○清水委員　　そうですね。

○佐久間部会長　なるほど。

○清水委員　　コンペティションの前まで非常に良い性能で、期待される手法ということです。それを上の図から拡大して持ってくると、このように実は 45 度からずれているということも見えます。

○佐久間部会長　例えとしては不適切かもしれませんが、試験対策でやり過ぎてヤマが外れたみたいなことになるわけですね。

○清水委員　　そうですね。

○佐久間部会長　逆に良いときはある意味では気を付けなくてはいけないことが有り得るのだということを意識しておく必要はあるのだけれども、それなりにそこそこのものであれば統計的に見るとそんな変なことは起きない、そういうことを言っているわけですね。

○清水委員　　はい、そうですね。ただこのケースが何をやっているのか、例えば先ほど具体的に心配したこういうことをやっているのか、それとも違うことなのかとか、そういうのを少しシミュレーションしながら実際に何をやるとどれだけバイアスが掛かるのかという具体例を出しながら議論をした上であまり心配しなくてよいとか、こういう注意点をもって開発する側も評価する側も行うべきとか、そういう知見を出していければと考え

ています。

○佐久間部会長　ありがとうございます。委員の先生方、何か質問があればお願いします。Gossmann の方法というのが、いわゆる評価結果を良くなる方向に持って行こうとするのだけれども、評価の評価関数というかこの場合で言うとそのアルゴリズム、AUC の差があまり大きくなったときは注意してその結果にノイズを入れてあげると、そこに対するオーバーフィットが防げる。そのようなイメージのアルゴリズムなのですね。

○清水委員　はい、おっしゃるとおりです。差が大きすぎるとテストデータの情報が漏えいして、テストデータに対するオーバーフィットというような書き方というのもあったと思いますが、そういうことが起こるのでノイズを加えなさいと。加えたものを使ってアルゴリズムを作っていけば、何度同じテストデータを使ってもフレッシュな場合と同じ性能がきちんと出てきて本当のことが分かりますということになっているようです。

○佐久間部会長　これは評価する立場からいくと気持ち悪いですね。非劣性という場合ならそこそこの結果なのだけれども、再学習しました、非常に良い結果になりました、認めてくださいといったときに気を付けなさいと言っていることなのかなと思えるので。

○清水委員　そうです。開発者は 80 何％という数値を出していても現実にそのシステムを適用すると 75％ぐらいしか性能がないとか、そういうような話になりますので、審査官はこういうバイアスが加わっていないか、審査だけではなくて開発者側もバイアスが加わっていないかというのは常に気にする必要はあると思いますが、過度にそれをやり過ぎると微小なバイアスに縛られて機器開発がうまくいかないというのはあまりよろしくない状況でもあるかもしれない。バランスが多分必要になります。

○佐久間部会長　分かりました。他に委員の先生方、ありますでしょうか。

○伊藤副部会長　教えていただきたいのですけれども、このテストデータを使った後にモデルを変えたり元に戻ってデータを漏洩するとかなりいろいろなことが起きるというのは分かるのですけれども、実際このような開発を進める上で豊富なデータセットを持てる人たちもたまにいるのかもしれませんけれども、持てない人が結構多いと思うのです。特に企業とか。どういう目安で具体的にここまでは OK ですとか、ここまではかなり危ないですとか、そういうデータのカットオフ値みたいな、例えばテストデータであればどこまでは学習用に持って振り返ってフィードバックしてもよいですよとか、そういう具体的なところがある程度あると審査なり薬事な

り開発する側も割とこの辺までは許容範囲なのだというのが分かるような気がするのですけれども、その辺の目安はあるのでしょうか。

○清水委員　とても重要なポイントだと思いますが、残念ながら私には明確なカットオフ値ということについて今ここでお答えすることはできませんし、恐らくそれは全部どんなCADの倍加によって結果が変わってしまう可能性がある話だと思っています。なかなか一般化しづらい話で、このCADの場合はここまで使っても大丈夫だったけれどもBという別のCADの場合に同じ結果が期待できるかどうかというのがなかなか分かりづらいもので、悩ましい問題だと思います。

ただ、だからと言って前に進むのを止めるのは良くないと思いますので、少しずつこういう研究、このKaggleのデータもそうですし、このGossmannの研究とか幾つか研究は進んで行くと思いますので、少しずつ蓄積されると思うのですが、それが十分なスピードで集まるかどうかというのはむしろ機器開発のほうがどんどん先行してしまう可能性は今あると思います。できる限りデータを集めていく必要があると思っています。

○佐久間部会長　他に、どうぞ。

○プログラム医療機器審査室主任専門員　一番最後のKaggleのスライドの所で解釈の仕方としてお伺いしたいのですけれども、今回１つのこのスライドの検討から言えるポイントとして、もしかしたら全体で見ればバイアスの影響というのはそんなに過敏になる必要はないのではないかという御示唆の１つだったかと思うのですけれども、これは例えば先生が５ページ目でしたか、Test用データセットのサイズを十分に大きくすればリスク低減ができるというコメントがあったと思うのですけれども、Kaggleの結果というのはデータセットが非常に大きいためにあまりバイアスの影響が見えなくなったという可能性はありますでしょうか。

○清水委員　これですね、例外として挙がっているのがテストの症例数が１桁、特に少ないような場合ですので、おっしゃるとおりこれも少ないのかどうかちょっとよく分からないのですが、少なくとも言えるのはもっとデータを多くしていけば、どんどん45度に近づいて行くと思うのです。ただそのスピードは問題ごとに変わってしまいますので、一概にはこうですということは言いづらいのですが。おっしゃるとおり増えれば45度にどんどん近づいて行くと思います。

○プログラム医療機器審査室主任専門員　ありがとうございます。あとこのスライドから

もう１つ、たくさんのモデルを全体で見ると 45 度の傾きに乗りやすくなるということかと思うのですけれども、我々としては目の前のモデルそれ自身がバイアスを持っているかというのは非常に難しい悩みどころです。ですのでもしかするとこのグラフの中の下側にいる１点のモデルが今、目の前の審査品目かもしれないと悩んでしまうところではあるのですけれども、これを見分ける方法というのは、やはり今のところなかなか難しいものでしょうか。

要は我々はバイアスがあるのか、ないのか、分からないのか３つに分けたときに、今のところほとんど分からないに落ちてしまっているから慎重な態度になってしまっているのが正直なところなのですけれども、何か見分ける方法、あるいは今日御示唆いただいた Gossmann の方法で２つのデータセットの差を見ることでこれはバイアスが乗っているとか判定できるとか、評価データからバイアスがあるかないかを見分ける方法は今のところありそうですか。

○清水委員　いや、決定的な方法はないと思います。もちろんフレッシュなデータを別途用意できればそれが一番良いのだと思うのですけれども、そうではない状況では今おっしゃられたような、テストデータの情報が漏洩しているかどうかを評価するこの式を使って、もしその企業が持っているトレーニングデータとかテストデータが使える場合にはそういうものを使った数値を出してくださいとか、そういうのはあり得るのかもしれないです。

ただこれはもう少し議論が必要な点がもう１つあると思っていて、先ほどテストデータに対する過学習、オーバーフィットの話だったのですが、ある機械学習のアルゴリズムがトレーニングデータにオーバーフィットした場合、簡単にこの差が大きくなってしまうのですね。それがトレーニングデータによる過学習なのか、テストデータにオーバーフィットした結果なのかというのを見破るためにはもう少し議論が必要なのではないかと思っていて、これを使うのは１つ考えられると思うのですけれども悩ましいところとも思っています。

○プログラム医療機器審査室主任専門員　ありがとうございます。そのデータから何か判定してこのモデルが少し怪しいからこっちのルート、このデータは大丈夫そうだからこの評価で大丈夫とか、そういうのができたら良いなと思います。ありがとうございます。

○清水委員　あともう１つ、設計の手順、プロセスをオープンにしていただいて、少

なくとも先ほど言ったような、こういうことはやっていませんねという確認は重要かと思います。決定的な証拠にはならないですけれども１つの状況証拠にはなると思いますし、これ以外にももしかすると幾つか考えられる可能性もありますので、そういうものを蓄積していってこういうリスクのある行動は取っていませんねという確認はしておくべきかと思います。

○プログラム医療機器審査室主任専門員　ありがとうございます。

○森委員　先ほどの Kaggle で試したものですが、すごく細かいことで恐縮なのですが、下のグラフのほうをよく見ると左から２つ目のグラフのレンジがものすごく狭いのです。

○清水委員　これなんですよね。

○森委員　はい。1%ぐらいの範囲で描いていて、他のところは 20～30%のレンジで描いていますよね。何かすごくこのグラフの出し方は印象をミスリードしているような気がします。上を見るととてもきれいじゃないですか。

○清水委員　多分この辺をぎゅっと拡大したのでしょうね。

○森委員　そうするとその辺が見ていてちょっと気になりました。あとこの論文の中で見ると、Kaggle のデータのほうの ID7634 というのが比較的データセットサイズが小さいですね。他の所が 20 何万件とか 200 万件ぐらいあって、このデータだけ 3,100。ただスピジーコンテストなので何やってるか分からないですけれども、ちょっとデータの性質が違うかもしれませんが。ただ 7634 が比較的データサイズが小さいのですけれども、またこの辺りがハッキングというか、中がどうなっているか知りたいと思いました。

○清水委員　n が小さいのは 7115 とかですか。

○森委員　7115 は。

○清水委員　5859。

○森委員　そうなんです。7115 は、サブミッションは 5859 なのです。7115 はそこだけは参加者がすごく少ないのです。5,800 人。他のところは 2 万 4,000 人参加していて、Public のデータの数が 5 万 3,000 で Private が 123 万です。7634 が Public が 3,100 で、プライベートが 15 万 5,000 なのです。だから何かちょっとまだ追いつけてないのか、パブリックのデータが小さい割には、以外に右から３つ目のデータがきれいだなと思っていて、何か考察とかがあったら教えてほしかったのです。

○清水委員　そこまでは深く考察できていません。n による影響ということですね。

○森委員　　n による影響がどれぐらいあるのかと思って見ていたのです。ありがとうございます。

○佐久間部会長　ありがとうございます。他にいかがでしょうか。よろしいですか。

○鎮西委員　　清水先生の話を途中からしか聞いていないのですが、例えば IDATEN に関して言うと、皆さんトレーニングデータを増やしていく話を一生懸命考えていたと思うのです。それによって性能が変わっていくという話なのですけれども、むしろテストデータを増やしていくという考え方は可能でしょうか。

○清水委員　　はい、学習のときに過学習等の学習が失敗してしまうリスクが少なければ、少数のデータから賢く学んで AI の性能を上げるという。

○鎮西委員　　皆さん今は基本的にそういう作り方ですよね。

○清水委員　　はい。それからテストデータを潤沢にして、評価のときに混入するこういうバイアスのリスクを下げる。そういうことができればよいと思います。

○鎮西委員　　そういう使い方、IDATEN で AI が賢くなることばかり考えている話が多いのだけれども、案外テストデータを増やしていくというのも意義があるかもしれませんね。

○清水委員　　はい。

○佐久間部会長　信頼性を上げるという意味ですね。

○鎮西委員　　そういうことです。

○佐久間部会長　他によろしいですか。

○岡崎プログラム医療機器審査室長　8 ページ目のスライドになるのですが、コンペティションの件です。テスト用のデータセットの議論ですね、2 種類の Test 用 DS の関係ということでお示しいただいたスライドの中で、コンペティションが行われていてそれぞれの成績が示されていたのがあったと思うのですけれども、このコンテストというか競技で、参加者がどういったことを手法としていたのかということを、例えば分析するなり明らかにすることで差分が大きくなるのか小さくなる要因の分析などというのは、この論文ではされているのでしょうか。もしそれがあるのだとすれば、先ほどおっしゃっていましたけれども、性善説に立ってこういうことはしないよね、ということで、チェックリスト的なものを示していくということになったときにその要素を提示していくことができるようにも思ったのですが、いかがでしょうか。

○清水委員　　重要な御指摘だと思います。私もまだ十分読み込んでいないのですが、

確か書いてあったのはやはり１人の開発者が１個のモデルだけで勝負するということはなくて、特に上位のほうは何度も何度もパラメーターを変えながら Public のデータセットに適用してその中から良いものを選ぶという、正に先ほどリスクがある例として紹介したようなことをしているようです。

多分おっしゃることは、何回やるとどのぐらいバイアスが掛かるとか、そういう細かいデータがもしあれば非常に有用だと思うのですが、そこまではないようです。もう少し他のデータ、論文にも当たってみないといけないと思うのですが、もしそういうものが手に入ると今、御質問の中で多分想定されているとは思うのですが、こういうことをするとこれだけバイアスが掛かってという統計データが得られますので、かなりいろいろなことが言えるようになるかと思います。残念ながら今のところはざっくりとしたようなことしか書いていないか、私が読み切れていないかで、十分なデータは持っておりません。

○岡崎プログラム医療機器審査室長　どうもありがとうございます。

＜検討の方針について＞

○佐久間部会長　よろしいですか。次の議題に移ります。報告書の検討項目と執筆分担の話に入りたいと思います。第１回の専門部会で説明したようにこの専門部会の目的は、審査上３つの論点に対して一定の考え方を示すことにあります。考え方までいかなくてもどのように気を付けなければならないかということを明らかにするということで、資料４を見ていただきますと、骨子(案)ということですが、多分３つの点と言っていたのは、データの再利用のあり方。今日の議論の所にもありますがそれと関連するところで数理的な視点からの考え方で、今日清水先生からお話いただいたようなことや、あるいは１と２に関連しますがデータ再利用の所で先ほど学習データの分布の話と評価データの分布の話の所をどう気を付けるか。また対応案として評価データの再利用により、改良プロセスと評価データを適切に分類しないと過学習ができてしまうので、それを例えば回避するためにどういうことに注意しながら見なければいけないのか。例えばこれは企業秘密になるので、多分これは PMDA が集めるしかないのかもしれませんが、どういうプロセスで良くしたいのかという話の所で最終的な結果、どうするかということです。

評価データと学習データの網羅性の話ですが、データを大きくするとい

うことから、今日、後で御議論いただきますが、先ほど森先生が御指摘くださったような、公共的なデータベース、AMED 事業等でもそういうものを集め始めてきています。AMED ですと大体数年で終わってしまうのですが、これは定期的にやったほうが良いとか、そういう提言を書くとか、そういうことになってくるのだと思います。

後で出てくると思いますが、医学サイドで使い方によっても随分気を付ける点は変わりますよねということがあるので、その辺りが出てくるのではないかということです。あと展望では個人情報保護の観点と倫理、そういうところが、骨子としてはあるかなと思いました。

執筆分担案みたいなものをそこに書いてみました。多分、序の主要論点のまとめのサマリーは私が書く形になると思います。数理的な所に関して言いますと、今日、清水先生からご講演があった内容についてお願いし、森先生もその辺りは御存じなので、特に数理サイドからという形になると思います。議論を受けて臨床領域のデータ収集という観点から、今日もメーカーから出てきましたが、やはり集めるのは大変だというのはどうしてもある。その中でリスクを抑え込みながらこの分野の発展を妨げないようにするにはどうしたらよいのか、正にレギュラトリーサイエンス的なのですが回答はないところでの妥協案を探していくという作業になると思いますが。これを実際に集めていらっしゃる先生方、使われている先生方の中で少し議論を進めていただきたいと思います。これは中岡先生には、これまでの研究成果を踏まえたところで、衛研という立場で、現在研究を進められている部分を踏まえて書いていただきます。あと倫理・個人情報保護法の観点については、実は今日コメントを頂いております。多くのデータベースを作られている先生方の中から、この点の懸念点、難しい点もあるということで、その辺りの現状ということもありますし、これは提言に近いことを書くことは実は許されている所だと思いますので、問題点にしていく所なのかなと考えておりました。

そういうことで先生方からこんな形になるだろうということを予測しながら、御意見を頂きました。お忙しい中、本当にまとめていただいて有り難いと思っております。これについて少しずつまた、一人一人御説明を頂ければと思います。森先生、バイアスについて整理されたことは、前半でもお話されたかと思いますが、資料に基づいてお話いただけませんか。

○森委員　　ササッと書いただけなので、どんなことがあるかということだけまとめ

ていただいてありがとうございます。バイアスについてはいろいろなものがあるのですが、以下に代表的なものを挙げさせていただきました。

結局、ある目的で利用するデータというものが母集団を正確に表しているかどうかというものと、あとは解析手法自体が持つバイアスというものと、あとはユーザーではなくて開発者が持っているバイアスが結構あるので、その人が自分が正しいと思っていても結構バイアスが掛かっているものはたくさんありますから、人に由来するものがあると思います。データが持つバイアスは、少しだけ整理したものとしては基本的には使用されるデータが母集団を厳密に表したものである必要があるわけです。一般には母集団全体を厳密に表現したものは不可能なので何らかのサンプリングによって得られる標本を用いることになるのですが、結局いろいろなサンプリングをやられるので、1、特定の病院だけで得られるデータ、これは特定病院でサンプリングされたデータと書き直しています。あとは 2、特定の機器で撮影されたデータです。これも結構バイアスが掛かっています。先ほどの富士フイルムの方からも意見が出てきたと思います。

あと 3、特定の病態・年齢、あるいは人種などから撮影されたデータということがあって、恐らく病気のものとかはよく話されていると思いますが、年齢や人種間の違いは結構あるので、こういったところも大きな要因のバイアスになってくるかと思います。

あと 4、手法開発に利用する際のデータ分割など、先ほどから清水先生からいろいろな例を提示していただいているかと思いますが、結局は手法開発で利用するデータはどうやって使うかとか、そういったところに生じてくるバイアスがあると思っています。

2 はあまり意識されていない場合も結構あるのですが、実際には特に人の目で見ても同じような画像にしか見えないのですが、ただ画像処理してみると全然違うということはよくあるのです。ですからこの機器バイアスというのも大きいのかなと思っています。

こういったことを防ぐにはやはり統計における母集団と不偏推定のような概念をきちんと理解して開発するということが一番私は大事であって、そのことを機器として承認を受けようとする場合には、先ほど清水先生にもお示しいただきましたが、どのようにやったかということをきちんと透明にしておくということが多分大事なのかなと思います。いずれにしてもこのバイアスをゼロにすることは現実的にはほぼ不可能なので、

透明性を確保するというところでバイアスが生じてもそれによっての影響を避けるというやり方を取るのが、実を言いますと最終的には現実的なのかなと思います。

2 番目に書いてあるのは、これはデータではなくアルゴリズムなどが持っているバイアスです。画像認識手法自体にはバイアスが当然あります。ある何とかのタスクに対して、○○しやすい傾向があるとか、××傾向があるというものはよくあります。特に学習型の手法の場合は 1 番目のデータバイアスによってかなり決まってくることになるのですが、もちろん実際にケア学習で使う手法の選択、あるいはモデルの形によっても変わってくると思いますが、そういったものによって生じてきます。あと画像処理手順によって構築された手法では、やはり手法が持っているバイアスが出てきます。あと人によるバイアスですが、先ほどからテストデータを選んでプライベートデータがどうのこうのと言っていますが、結局、人によってデータというのは取得されて選択されますので、最終的には人によるバイアスがどうしても気が付かないバイアスというものがあるかと思います。

そういったことはきちんとした教育をするとか、きちんと理解しているかどうかを確認することが大事かと思います。

ここに書かなかったのですが、本当に気が付かないうちのバイアス、無意識バイアスで、最近ジェンダーの関係でよく出てくると思いますが、実を言いますとこういったものについては開発者のバイアスが結構掛かったりすることもあります。

結構こういった人によるバイアスというのはあるのですが、これは過去に遡っていろいろ検証しないと分からないこともあるのでしょうから、なかなか難しいのかなと思います。その他、幾つかのバイアスが出てくると思いますので、報告書作成時までには抜けている所を記していきたいと思っております。

○佐久間部会長　ありがとうございます。何か委員の先生方から御質問等ありますか。バイアスをいくつかに分類してどういうところにありそうだということを明らかにして、先ほど使用者の先入観に関する話がありましたが、その辺りを明確にしていくことで多分リスク低減が図れるのではないかと思います。

○森委員　そうですね。1 番は完全にいわゆる確率統計の教科書で出てくる基本的な所が出てきて、あとはケア学習で拡張していくということだと思って

います。3 番は完全に社会学的な話ですね。2 番は多分工学的な問題だと思います。

○佐久間部会長　先ほど先生が前半でおっしゃっていた、AI のシステムを使って慣れているという話で、例えば違う AI を使ったときに中を変えてしまう、あるいはもともとそうなるはずだというところで判断していたところが、実はそういう AI を使うことによって違う判断に至ってしまう。そのおそれですよね。これはかつて SaMD の話をしたときに、いわゆる電子カルテのシステムで、禁忌の薬を処方されないようにする機能をある病院でカスタマイズして、そこで働いていたお医者さんが違う病院に行ってそこでカスタマイズされていない電子カルテを使って薬を出してしまったケースがあるのです。ここまで極端ではないですが、そういう可能性はあるということを今日御指摘されたのかなと思います。

○森委員　そのとおりです。結局どこまでを評価対象とするのかによると思うのです。機械が出してくる出力までがいろいろな評価対象になるのか、それを人が使って最終的に行われる診断が評価対象になるかということですよね。

例えば車のナビゲーションでもいろいろと警告が出てくるけれども、結局みんなこれは関係ないから無視しようとだんだんなってきますよね。それと同じようなことが多分起きてくるので、これは実を言いますと機器認証の評価にされるのかどうか分かりませんが、意外に大事だと思っています。

○佐久間部会長　その辺りは多分使い方のガイドラインみたいな。

○森委員　そうですね。

○佐久間部会長　他の先生方、何かありますか。

○鎮西委員　こういうことを言ってはいけないのかもしれませんが、そもそもバイアスがあることが悪なのではないというところをきちんと全面に出したほうが良いのかなと思います。

○森委員　そうです。私もそう思います。

○鎮西委員　無にできないということもありますし、逆にそもそものこの種の機械学習のアルゴリズムというのは良い成績が出るように一生懸命調整して作るものですから、これは普通は統計学的にはやってはいけないことをむしろ積極的にやってそれによって良い結果を得るというものだと思っています。

それでいきますとバイアスがあるということは前提として、出発点とし

て扱うべきなのかなというのが１つです。もう一つは会社が最終的になぜ市販後学習をやるかを考えますと、例えば適応症例が増える、あるいはライバルがいてライバルよりも良い正診率を得るとか、それぐらいしか思いつかないと思うのです。より良い正診率を得ることに関しては今日やってきた議論でよいようにかなりいろいろなところを尽くせていると思うのですが、例えば適応拡大のことを考えたときにどういうことを考えるべきでしょうか。今回はもう扱わないことにしますか。

○森委員　　佐久間先生、どうぞ。

○佐久間部会長　扱わざるを得ないんじゃないですかね。それは違う問題だということを指摘して、問題の特性みたいなことを明確にしておくだけでも実はよいかもしれない。

○鎮西委員　　多分会社になったつもりで考えますと、適応を決めるときにどこからどこまでの範囲にするというのは最初にデザインしたものがあって、それに対してデータを一生懸命集めていったところこれはそこそこの成績でいけるかも、ただしボーダーの部分に関してはどうなるかよく分からないということでそこをばっさり切るとか、そういうことをやっているのではないかと思うのです。それがだんだんデータが集まってくることによって、例えば９割という正診率を目標にしていてそこは何とか統計的にそこまでの成績が出そうだということになったからその部分も含めて適応拡大の申請したいですという形になるのかなと思うのです。その場合であれば今日やってきた議論で正診率の向上とか、そういう話で済むのかもしれない。

ただ一方で今まで集めていなかったデータも含めて集めるようにして、その結果入れるようになりましたという例に関してはどうすればよいのかなというのが、今日の話とまた別の検討が必要なのかもしれません。

○佐久間部会長　他にありますか。データベースの作成時の課題ということで出していただきました伊藤先生から、また御説明いただけますか。

○伊藤副部会長　AMED の事業で、内視鏡外科手術のデータベースを集めてきました。本年度の 3 月 31 日までやった事業です。胃がん、大腸がん、胆石、前立腺がんで約 4,000 近い動画情報のデータベースができているというところです。これは当初書いたときにうまく整理できていなくて申し訳ないのですが、結局やりながら幾つかの壁というのにぶち当たってきたというのが事実です。大きく分けますと、データを準備するまでと、それの扱い方法、個人情報、その 3 つの柱かと思っています。

五月雨式にここに書いたのは、動画の権利、帰属の曖昧さと書いたのですが、実は動画というのは誰のものなのかという部分で、手術をした医者なのかその病院の院長なのか患者さんなのかというところで、一応我々の同意書には全ての方にサインを頂くことにはなっているのですが、この辺の曖昧さがまだある。

2 番目として、大規模データベース、手術の動画のデータベースを扱うときに、通常、匿名加工情報で利用するというところですが、加工情報で使えるものもあるのかどうかということも我々の中で議論をしてきました。

データベースと言いますが、内視鏡外科の手術というのは動画の形式というのが結構施設によってばらばらで、カメラメーカーもオリンパス、ストルツなど幾つかあるのです。録画システムも保存形式などありまして、結局 FPS が合わせたり、解像度を拡張したり、BPS を合わせたりといったことがあります。

4 番目の動画の匿名化の方法に関しては、体外に映り込んだ人の顔を黒塗りするだけで足りるかと書いてありますが、そういった問題があるということです。いつもこういったところで引っ掛かったりいろいろ議論になるのは患者同意のサインの利用目的の明確化の難しさというか、内視鏡外科学会にも通じて技術認定の手術ビデオというものを全て患者さんの同意の下でこちらのデータベースに頂いているのです。これに関して言いますと、今は産業利用、教育・研究を含めた患者さんの 2 次利用に対して承諾を頂くという形でやっているというところです。

あとお金の問題だったり、サステナブルにこういうデータベース、AMED の切れ目が縁の切れ目ではないですが、金の切れ目でこれが終わってしまってはいけないので、これをいかに持続的に運営してキープしていくかという部分において、そこの課題もあります。我々においては国立がんセンターの認定ベンチャーというのを作って、そこにおいてそういったものを維持するような方策を今作っているところです。

○佐久間部会長　ありがとうございます。何か先生方からありますか。やはり産業界の方もおっしゃっていましたが、評価のデータとか、そういうことをするとか、返せるデータにするといったときに結構重要ですから、その辺りの目的まで含めてどうやってやるのかは課題ですね。

○伊藤副部会長　あと 1 点ですがやはり手術動画は、先ほどの診断みたいな CADe、CADx みたいなものとまた違う将来的な使い方なのですね。その出口において

若干毛色が違うなというところも、審査する側の人もそうですし開発する側もみんな理解しながら進めていかないと同じ議論ではなかなか立ち行きにくいところもあります。

○佐久間部会長　笹野先生、次に心電図データベースを作られるということで、これまでの議論を聞かれて感じられたこともあると思うのですが。

○笹野委員　心電図とか、光電脈波なども含めた生体データのデータベースということかと思って、ここに書かせていただきました。私たちは AMED で 3 月で終わったのですが、心電図、通常の 12 誘導心電図ですが約 2,700 例集めて発作性心房細動の人を予測するというアルゴリズム、AI のモデルを作ってきました。そこで出てきた問題点をこちらに列挙しております。

1 番の external validation のデータセットですが、external validation が必要か必要ではないかと言えば当然必要ですが、それを例えば私たちは他施設でデータを収集してただ機器は統一するという形でやったのですが、validation の施設は 1 施設任意に選べばよいのか、汎用性を担保できるような validation の施設、あるいはサンプルというのを選択して用意しておくべきなのかというようなことは最初に決めておかなかった。今でもまだ結論としては出ていなくて、今、新たに収集している段階です。

汎用性の問題にも関連すると思いますが、validation をどのように評価するかということ。私たちのモデルでは最終的には健康診断などでの応用を考えています。そうしますと確実な validation というのは健診データを用いた前向き試験ということになるのですが、それを行うのは現実的には非常に困難で、有病率の低さなどもあって、それはどこまで許容されるのかということをある程度示していただくのがよいのではないかと思いました。

2 番目ですがこれもやってみて後から分かったのですが、臨床データをそのまま使うということであれば、そのまま AI、プログラム機器として使えるのですが、例えば各病院で心電計のメーカーは違いますので統一した機器でやろうということで私たちは心電計を各病院に持って行って記録をしたのですが、それは研究として取ったデータであると。なので診療として取ったデータではないので、研究データをそのまま使うのはよろしくない。使うのであれば GCP 省令準拠のグレードで取らなければいけないということが後から出てきました。

ただそれを多数の症例で行うというのは、最初にそれを決めるのはなか

なか難しいかという問題はあります。そこもどこまでのグレードが必要なのかというのは、どこかで示す必要があると思います。

3 番目、教師データが医師の診断で確定できない場合。今日、富士フイルムの成行さんの話でも例えば正解データが 100%正しいかという話がありましたが、私たちのほうでは陽性コントロールが 100%正しいかどうかが評価できないという問題があります。それは発作性の不整脈を診断する場合に、非発作時を見ても誰も分からない。専門医が診ても分かりませんし従来の方法では分からないということになりまして、そこのコントロールをどうやって担保するのか、究極的には生まれてから今まで心電図をずっと取り続けている人以外はコントロールになり得ないということがあって、どれぐらい記録をしていればよいのか。そのようなことがなかなか難しい問題で、これは必ず議論になるところなのですが。このようなところの取扱いの基準などもある程度示すのがよいのかなと思いました。

ここには書いてありませんが、心電図以外について。12 誘導心電図などはある程度汎用性を持ったものが作れるとは思うのですが、それ以外の医療機器承認を受けたウエラブル機器みたいなもので脈波センサーなどからプログラムを作っていく場合には、やはり機器間のバリエーションが非常に大きくて、汎用プログラム AI ではなくて、機器特有のもので作っていくしかないのかなと今思っています。それは今日の議論を拝聴していて思いました。

○佐久間部会長　ありがとうございます。やはり心電図固有のバイタルサイン、それとデータの取り方に対する論点を示されたのだと思いました。何か先生方から御質問はありますか。田中先生、次お願いします。

○田中委員　書いてあるとおりです。実はもう少し長い文書を書いたのですが、数行にと言われたのでこのように書きました。内視鏡検査というのはワンセクションで大体 40 枚ぐらい写真を撮るわけです。食道、胃、十二指腸、上部だとすると 3 臓器にわたるので、やはりそれぞれのテキストデータが標準化された記載がないと後で使いものにならないというところがありますので、やはりテキストデータをバインディングして持つことが重要であると。かつ、病変を囲むアノテーションなどがありますので、それを XML 形式でこういう規定でというのを学会では作っているというところがあります。

ただこれは先ほど伊藤先生もおっしゃってましたが、研究なのかそれ

とも商用転用なのかで全然同意のあり方が違うわけです。我々からしますと臨床研究法のように介入のある前向きでやるわけではないので、後向きの非管理のものに関してどこまで何をやっていけばよいのか。今、我々としてはアルゴリズムを作るまでは研究でやって、validation に関しては個別同意を取るというような形でやっていますが、この辺りの整理というのはまだできていないのではないかという思いもあります。その辺は大きな課題かなと思っています。

○佐久間部会長　何か御質問はありますか。テキストデータに関して、この映像ではこういう所見があるというテキストデータはしっかりタグ付けられた形で入っていて、かつまたそこでどこを見ているかというアノテーションもあるということですよね。

○田中委員　そういうことですね。今は癌か、癌ではないかとか、癌を見つけるとかそういう話になっているのですが、我々が普段診断名として用いる用語というのは胃だけでも最低でも 40 種類ぐらいあるのです。それをマルチパーパスの AI を作っていこうと思うときに、やはりしっかりと構造化されたデータのバインディングが必須になると考えています。

○伊藤副部会長　先ほど validation までは研究で、そこからはもう一回取り直してテストをするというニュアンスでおっしゃったのですが、結局モデルを作るところまでの動画なり画像情報をモデル作った後、商品化はできるのですか。

○田中委員　いえ。validation を個別同意で取って、研究外です。個人情報保護法にのっとって validation を取ってそれからでないと商品化できない。

○伊藤副部会長　もう一回そこから同じことをやってアノテーションして、モデルを作り直すということですね。

○田中委員　そういうことですね。完全に前向きで、ピュアなというか、ナイーブな症例に関して AI の性能を判断するということが必要なのではないかというのが、今の私の判断です。それが正しいかどうかが私もよく分かっていないのですが。

○伊藤副部会長　分かりました。ありがとうございます。

○田中委員　ありがとうございます。

○佐々木委員　病理学会でもデータの扱いで非常に困ることがありまして、結局後ろ向きで集めた数十万というデータを、商業ベースで企業などに我々AI を開発するよりも、AI の技術、知識を持った企業さんに参画していただき最終的には保健診療所で使えるようなプログラムを開発するときに、商業

ベースという言葉がいつもくっ付いてしまうのです。

後ろ向きで集めたデータを今、田中先生の話にもありましたが、もう一回患者さんの同意を取るのではなくて、結局最終的に良いように使うということで患者さんも書いていくものだと思うのですが、企業さんなどがある程度商業ベースでそれを利用できるようにする仕組みというのは難しいものですか。データを集めるときも商業ベースだと提供してくれないというところが結構あって非常に困っているのですが、どのようにしたらよいのかということが１つです。

○佐久間部会長　科学委員会の範囲を超えてしまっている議論かもしれませんが、科学的にやろうとするとここがやはり社会システムに関わる障壁になっているという指摘はできると思うので、書き方を注意していくことで何らかの形に残したいなとは、部会長としては思っているのですが、うまい表現を考えなければいけないと思っています。逆にこういう製品の評価、科学的に行うためにやはりこれは必要なのだけれども現状そういうところは社会的な受容性が低いために実施しにくいとか、そういうことについては少しその観点から書いたらよいのかなと思っていますが、いかがですか。

○伊藤副部会長　今言われた話は、多分皆さん同じことを思うと思うのです。ただ我々は他のデータベースをやられていた先輩方を見て、最初から産業利用のオプトインをしてやろうというので取ってきたものなのでそこまでできるのですが、ただ先生がやられているようなレトロでも、マスキングしたデータとか、モデルとか、その辺は匿名加工情報として扱えるのであれば、それを使って産業利用してもよいのではないかという議論はないのですか。

○佐々木委員　厳密に言いますと、おそらく次の項目にも関係あると思うのですが、我々病理学会で集めていたデータはいわゆる匿名化された情報で、いわゆる個人情報保護委員会が定めたような匿名加工情報の手順に従って作った情報ではないのです。ですので実は匿名加工情報という呼び方には当たらなくて、それで使えないですよというような指摘も出ています。

○伊藤副部会長　生ではなくてその後に結局マスクしてアノテーションしたら、マスク動画だって誰が見ても分からないではないですか。その後のモデルも結局、それは誰のものか分からず、匿名加工化されているのに近いのではないかと我々は議論をよくするのですが、そこに対する明確な答えを持っている人は誰もいないなというところがあります。そういったところもど

なたか教えていただければ有り難いのですが。

○佐々木委員　前に弁護士さんに聞いたときには匿名加工情報と匿名化された情報というのは違うという話をされて、匿名加工情報というのはいわゆる法にのっとって手順どおりにやったもので、病理学会がやったものは匿名化された情報という学術研究に用いるためのデータ資料になりますよと指摘されました。

○佐久間部会長　もし中田先生、何かコメントがあれば。今、答えを出せと言っているわけではありません。これはここでできるかどうか微妙な問題も含めて何かありますか。

○中田(はる佳)委員　今、先生方御指摘くださったように、後ろ向きで集めたデータを商業利用するにはどうすればよいのかというのはいろいろな所で指摘されている課題だと思いますので、そのことについて先ほど部会長から書き方を工夫するという御発言がありましたが、それについて検討するということは非常に有用になるのではないかと思いました。

もう１つは今、先生方の御議論を伺っていて、指針の中で使われている匿名化された情報とか、あと個人情報保護法の中で使われている匿名加工情報という用語の整理を、どこかでできたらよいのではないかと思いました。

○佐久間部会長　ありがとうございました。その辺りが科学委員会的にはできるところかなと思っています。先生方の問題意識はよく分かりますので、そこをうまい形で表現できるように工夫したいと思います。

○佐々木委員　この４月１日での改正個人情報保護法が施行になってからかなり状況が変わっていると思いますので、また少し整理ができればよいかと思います。

○佐久間部会長　私の不手際で議論したと言えば議論したのですが。

○中岡副部会長　今の御議論は随分昔からあった話で、匿名化した情報というか集めた情報をどこまで開発で使えるか、研究の範囲でしか使えないかというところの切り分けは大分前から多分問題視されていたのですが、今、言われたように、誰も実はなかなかこれに対して答えが出ていないと。個人情報保護法を改正されて仮名情報はできましたが、実は仮名情報というのは何かという定義がきちんとされていないところが多分にあって、そういうところに対する提言という形であれば何かしらできるのではないかと思いました。

この問題を語るときにもう一個だけ忘れてはならないのは、これまで実

際にオプトインという情報を集めてきてきちんと法に則って集めてきて開発をされてきた業者ですね。結局これもまた曖昧なままに後ろ向きのデータを開発でも使えますよとしてしまいますと、今まで正直にやってきた方たちの反発もあるかと思うので、その辺も考えておくとよいかと思いました。

○佐久間部会長　御議論ありがとうございました。本日の議事は以上ですが、特に先生方からありませんか。事務局に戻させていただきます。

＜その他＞

○事務局（渕岡先端技術評価業務調整役）　次回の専門部会は、12 月 5 日(月)15 時から 17 時の開催を予定しております。詳細等については、追って御連絡いたします。

また、10 月 24 日(月)15 時から 17 時には、第 2 回のワーキンググループを予定しております。御参加いただく先生は、伊藤副部会長、佐久間部会長、佐々木委員、笹野委員、清水委員、田中委員、鎮西委員、中岡副部会長、中田典生委員、森健策委員です。

＜閉会＞

○佐久間部会長　またワーキンググループで今日頂いた議論を少し揉んで、方向性を定めていきたいと思います。本日の専門部会はここまでとさせていただきます。先生方、どうもありがとうございました。